Beyond an Anthropomorphic Template

Dr John Elliott
School of Computing, Creative Technology & Engineering
Leeds Beckett University
Leeds, LS6 3HE, England, UK
J.Elliott@leedsbeckett.ac.uk

Abstract

In our endeavours to explore all possible forms that non-terrestrial communication may encompass, eventually we must throw off our anthropomorphic bias and investigate the implications of post-biological intelligence on SETI search strategies.

In the event a candidate signal is detected, our initial categorization and assessment will focus on analyzing its comprising constructs, to ascertain whether information content is present; a fundamental signature of intelligence. To ensure our systems are capable of encompassing such intelligent communicators, we need to investigate both the contrasts and similarities of such non-biological communication and how this extends the known spectrum.

In this paper, we begin to investigate the likely signatures and contrasting structures such non-biological communicators may present to us, across a range of known machine communication phenomena, and discuss how such contrasting forms of information exchange can aid, extend and refine our detection and decipherment capabilities.

Introduction

In Shostak's paper "What ET will look like and why should we care" (2010), he highlights "our anthropomorphic bias about extraterrestrials" and the implications of post-biological intelligence on SETI search strategies. Although the rationales behind searches to detect non-biological sentience are not our concern in this paper, our remit is to investigate the likely signatures and contrasting structures such non-biological communicators may present.

Above all else, we use communication [language] to convey information to someone or something else. Whether the conduit for this information is vocalised, written, or gestured, our purpose remains the same. In order that the message is understood, we use a shared 'code-book' of established [agreed] abstractions [symbols, sounds, movements] to represent the meaning of our 'message, so the information is understood by the recipient.

Like any system, shortcuts can be made, when the resending of the message is 'cheap' and quick, or the context removes any possible ambiguity. Correction of incorrectly interpreted information or repetition of information, due to loss or 'damage' of comprising segments, can be costly; so, language evolves to negate (or at least reduce) such overheads, establishing rules and redundancies. Language is also structured according to the abilities [cognition; vocal dexterity] of the system users, to facilitate efficiency.

The natural communication system I describe is not a theoretic optimum but a highly efficient compromise. Humans typically cannot retain more than nine pieces of information at any one time, in their short term memory. Due to this limitation in our processing abilities, the communication system we have evolved uses inbuilt mechanisms [clauses, phrases] to structure information in suitable chunks for processing. Nevertheless, the system is highly efficient and dynamic, having the capacity for infinity variety.

The 'forces' at work in evolving such a system of communication [language] is also shaped by the requirements of speaker and hearer [recipient]. In this, a form of reciprocal altruism is embodied, which follows a principle of least effort: compromising between the speaker's need to efficiently communicate the Information and the hearer's need to receive and understand the message, unambiguously.

There is no common semantic assignment or common syntax: the veneer of communication. However, English is representative of a typical human language, in respect of its underlying structural signatures, for alphabetic [phonetic] based encoding, as well as representative of human language's entropic [information theoretic] signatures. It is the veneer of the sounds we utter and words we arbitrarily choose to assign semantic values and, the morphology system to 'glue' this together, that lends us to perceive the way we communicate differ significantly. When we strip this veneer away and look at the underlying structure, such as phrase chunking, where cognitive constraints operate, and internal structure of conditional probability, we then see the human language machine's [human brain's] common 'footprint'(Elliott, 2002).

Our observations are admittedly from a single source: our own planet. However, extensive analysis of a wide variety of human variants of language has shown that they all adhere to the same underlying structures, dictated by the aforementioned constraints. It has also been possible to analyse the communication of a variety of other species, which arguably constitute 'aliens' on our own planet; for which results show the same 'forces' at work in their structure: results that support the principles of communication being 'universal' and therefore identifiable, where discovered.

It is postulated, when communicating across the vast reaches of space written [text] communication is the most likely and is therefore the focus of analysis, for this paper.

Beyond the Anthropomorphic Mold
Nevertheless, in addition to the constructs of the natural language we use, with its evolved efficiency, range and flexibility, there are known alternative methods for representing

information and knowledge. Given the possibility that such methods may be adopted, especially by non-biological forms, we need to look at the signatures such systems would present, to ascertain whether they are readily distinguishable.

One of the principle candidates, amongst these alternative methods for communicating information, which has been considered as viable for interstellar communication, is logic. Here mathematics meets semantics, in formal constructs, which convey information and semantic relationships in precise and explicit terms. Admittedly, I cannot recall any human actually using this method of communication, as a preference to natural language, but its potential for precisely [unambiguously] encoding and relaying information must make it a possible conduit for interstellar and remote intelligent inter-species communication. See LinCos section below.

Would a machine construct [evolve] a communication system based on logic or an optimised form of natural language encoding [no redundancy - 100% pattern utilisation]? We will look at examples of known constructs in the machine [assembly] code lexicon and evolved robot communication from recent applied research, where the arbitrary pairing of a linguistic label to its assigned meaning are agreed by the robot community, without human intervention: The Lingodroids project (University of Queensland, Australia and Queensland University of Technology).

(Robot) Silicon chat

Lingodroids are robots, which use an onboard camera, sonar and a laser range-finder, to map the space around them (Schulz et al, 2011). This language, which sounds similar to the tones on a phone, is 'spoken' aloud by using a microphone and speaker. Experiments conducted in this project are a useful insight into how machine intelligence may develop communication.

The communication they use is not a typical computer [programming] language, but more of a human language. These words have been 'invented' by the robots themselves, using a variety of games to establish correlations between specific words and places, directions, and distances. And this includes teaching themselves brand new words for different lengths of time.

Although the Lingodroids described demonstrate the vital role played by communication for any task requiring more than one individual, the current state of evolution is nowhere near the complexity of a mature (fully developed) language, which can embed information (clauses, phrases). Nevertheless, the basic concepts [building blocks] of language are developing, akin to those seen in animals, where relaying such information is vital for survival: where did you find the food? Where is the danger? Finding a mate, etc.

Examples of the Lingodroid's vocabulary show consistent use of short 4 letter words, to represent place names, distances and periods of time;

significantly less word length variation than in Human language. This supports the reasoning that machine language will develop their vocabulary, to explore all permutations within available variables, for maximum efficiency, as physiological and cognitive constraints will be negligible, in comparison to biologically based life forms.

> Geographical location examples: yifi, kiyi, gige, mira, xala, soqe, sihu, juhe, rije, pize, tuto, kopo, heto, qoze, yaro, zuce, xapo, zuya, fili.
>
> Distance measurement examples: puga, puru, vupe. duka, ropi, puga, huzu, hiza, kobu, bula,
>
> Temporal examples: kafi, puni, fohu, qija, fedi, tofe

Unlike human discourse, where a given language only explores a subset of the phonetic space, typically, a given language will only use on average 50% of bigram (two letter) and less than 20% of trigram (three letter) possible combinations (Elliott, 2011). However, robot [Lingodroid] vocalisation is able to explore its entire vocal space and requires no inbuilt redundancy, akin to dialing phone numbers. So, optimisation of ngram usage - for a given word length - is both possible and desirable, for purposes of efficiency.

Entropic Signature (see Glossary)
Unsurprisingly, entropic measures for information content are similar across samples, for human, robot language, Logic [LinCos] and machine code, as they need to contain similar complex and information rich content. However, with the caveat that samples of Lingodroid discourse are currently limited to a relatively small corpus, results from calculating and comparing these samples do show a similar but distinguishable range of relative entropic values across the range analysed: H0 to H5. See Figure 1.

As all samples demonstrate evidence of internal structure, by virtue of their entropic values decreasing as dependency increases (across the range analysed), no significant difference, immediately serves as classification indicators for categorisation. However, it is apparent that the internal structure of natural language is distinct from the other two samples (Computer code; Lingdroid), which demonstrate very similar relative values [slopes]. See fig 1. Note: This phenomenon is also true for the logic language LinCos. It is also postulated that, as the sample size of the Lingdroid data set increases and diversifies (probably within its continuing use of 4 digit [tone] sequences, the higher order entropic measures will tend towards a -1 slope.

Semantic Encoding
Any communication system requires the information to be transmitted unambiguously across its channel. A distinguishable atomic unit within such a written or spoken system, which is suitable for communication across distance and time, is categorized as a word. Varying devises are built into these words, to further aid both semantic disambiguation and coping strategies for noisy channels; typically, this includes additional information, such as case endings and redundancy.

Given this, analysis of the word length distributions across the three data types (Computer code, Robot communication, and Human language) discussed in this paper, were compared, to ascertain any significant variation, as an aid to classification. See Figure 2.

As previously mentioned, human communication (and evidence indicates all biological communicators) only use a subset of possible combinations from the comprising auditory and - in humans - corresponding textual lexicon. It is believed this is to facilitate disambiguation for audio reception. This subset will vary between languages but the 'area' each subset covers is of a similar percentage of all possibilities.

Given this, by analysing and comparing the ngram profiles of languages (measuring the percentage of bi & trigrams: 2 and 3 letter combinations in their written form), it is postulated that the Lingdroid lexicon will tend towards 100% coverage, in contrast to biological [human] systems, which only utilise a subset of possibilities; a phenomenon that in part is probably symptomatic of our physiological constraints. An observation likely to assist distinguish between non-biological and biological communication, as there is a typical statistical relationship across the symbol [phonetic] set: in Human language, the typical percentage of bigram and trigram letter combinations are 60% and 10%, respectively: 6:1 typical ratio (Elliott, 2002).

Possible combinations, based on 26 letter alphabet, which is, of course, an assumption, but the following demonstrates a system's potential, where constraints are removed:

Bigrams (2 letter words) = 676
Trigrams (3 letter words) = 17, 576
Tetragrams (4 letter words) = 456,976
Pentagrams (5 letters) = 11, 881, 376

Therefore, word length where no constraints exist for vocalisation have no need to exceed 5 tokens in length, to generate a lexicon capable of exceeding all requirements to communicate. A simple check sum at the end of transmission (such as with Hamming codes) will assure no errors have occurred.

Logical devices [operators] will need to be fully developed, to enable full descriptions of context and relationships. In these, it may be the case that such 'words' are even shorter, for efficiency and differentiation. In effect, function and content words, where the function words are logical operators such as "and", "or" and "not".

In addition to this, the evidence of morphology and concatenation of patterns are useful indicators of an evolved system that has economy, recursion and embedding. For a biological system, like our own, where cognition is a key limiting shaper of processing information, such components are highly likely to be evident, as they greatly assist learning and cognition. However, current (limited) evidence is indicating that when the cognitive 'breaks' are off, the need and utilisation of such mechanisms are much less likely; and in the case of Lingdroids, completely absent.

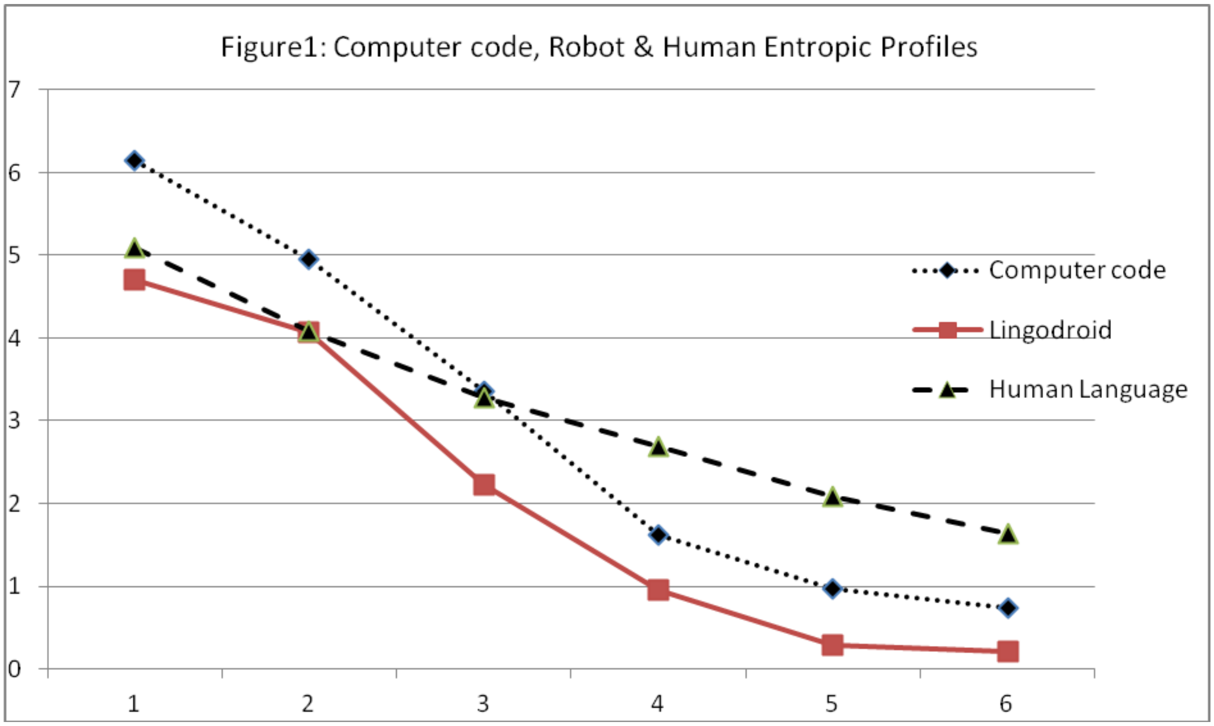

Figure 1: X axis - Entropic Order (as value increases, so does conditional probability);
Y axis - Entropic Value of sample, at the given entropic metric. H1: H6.

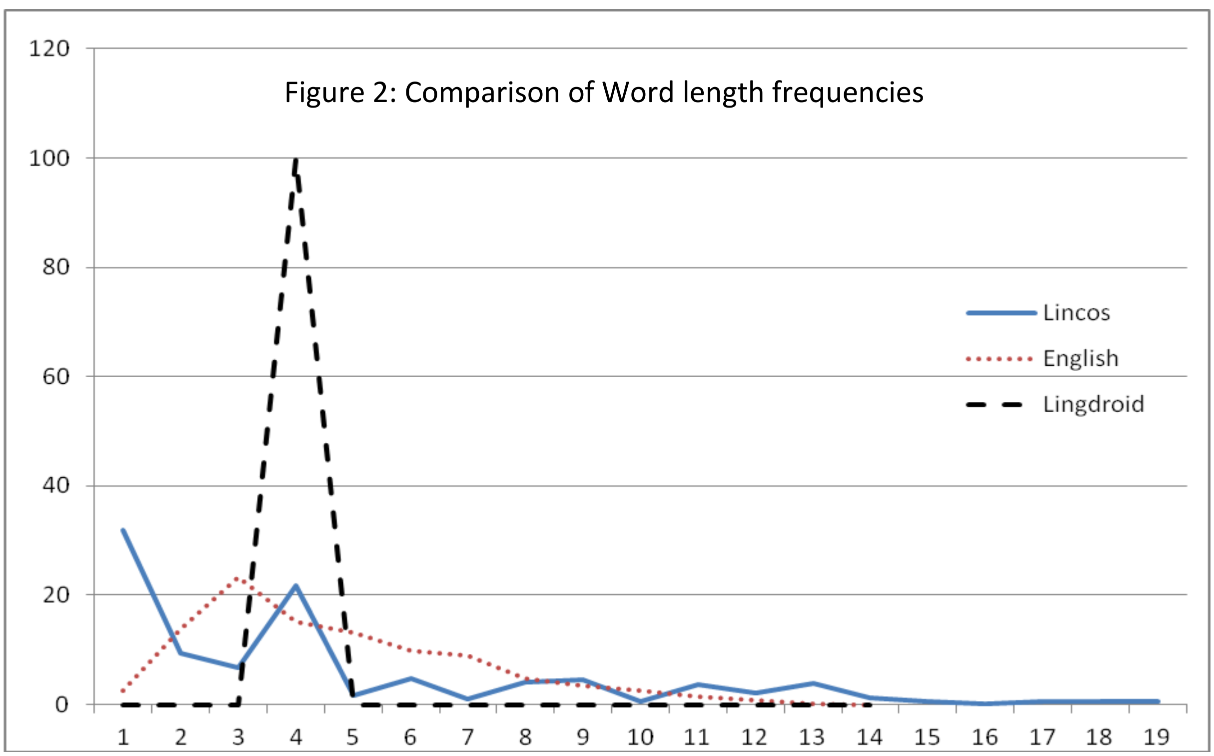

Figure 2: X axis - Number of symbols in word;
Y axis - Percentage frequency of word length, across samples analysed.

Examples of morphological mechanisms in human [biological] language can be observed through statistically significant increases in a given grouping of a particular affix, far exceeding what would occur through random behaviour. In English .the following two constraint pairs demonstrate examples of this phenomenon: ‘which-the’ reveals 80% of suffixes end in ‘ed’; ‘in-the’ reveals results of 59% of suffixes ending with ‘ing’ (Elliott, 2003). Although such grammatical devices are subject to individual language rules, it is a recurring pattern phenomenon that such groupings occur.

LinCos: Lingua Cosmica
As previous referenced in the Entropic Signatures section, LinCos is a logic-based language and is therefore a candidate for communication beyond the anthropomorphic template. This logic-based artificial language was first described in 1960 by Hans Freudenthal in his book Lincos: Design of a Language for Cosmic Intercourse, Part 1 (Freudenthal, 1960), and later extended by exponents such as Alexander Ollongren (Ollongren, 2011). I have included a section on this particular system, as it typifies communication through logic, as an alternative to natural language.

In figure 2, it can be seen that the frequency profile of LinCos, compared to human language [English] displays atypical variations, against a consistent profile that is true across alphabetic systems: notably, a spike at 4 letter words. Where natural language follows consistent frequency profiles, for efficiencies of effort and cognition, logic is an extension of such a system but is rooted in formal constructs, variables and constants, all of which are encoded in arbitrarily chosen, yet consistently short words; or assigned to single letters. A dynamic seemingly chosen for efficiency of encoding, as normal communication requirements, such as redundancy, do not apply. LinCos is also clearly distinguishable at the phrase analysis level, where it is also evident cognitive constraints do not apply (Elliott, 2011); Higher order Entropic profiles are also

The logic constructs, which underpins such a language are dependent on correct semantic interpretation, usually via convention or explicit assignment. Such assumptions imply compressed knowledge of more detailed descriptions of the world and therefore are unlikely to be used in isolation for communicating complete reasoning and understanding.

It is postulated that such a logical construct is a product of an intelligence, with a mature natural language, and not one that is evolved as an alternative. Either that or the communication can only remain as an inefficient terse pidgin-like language for communication of simple concepts and descriptions.

Nevertheless, although not 'chosen' by the Lingdroids, it is still a possible candidate and one potentially used to assist communication of explicit relationships, between Worlds where there is no established shared code book.

Conclusions

Communication, no matter how 'alien' needs to be received in a form that conveys the content generated by the author, which is unambiguous and complete to the receiver. Or, at least in a lossy state that maintains the content decipherable in its intended form. This inbuilt robustness can enable a receiver to correctly interpret information, which has suffered significant loss and we often construct games to demonstrate this; we will even automatically correct any typos and omitted words, as long as the context enables us, in real time.

The mechanism [system] by which the information is carried and semantically encoded can and will vary, dependent on the medium and constraints of the intelligence involved. In the case of non-biological communicators, such as Lingdroids, the mechanisms by which information is relayed have demonstrated the potential to explore the encoding of information within a more concise and fixed length, as the need for inbuilt redundancy, recurring patterns, to address cognitive limitations, is no longer a consideration.

It is possible non-biological systems could begin encoding their utterances [semantic assignments] at 1, 2 & 3 letter word combinations but these would soon be exhausted (especially 1 and 2 letter words) and latter development of function words, as logical operators, which develop post assignment of naming things, attributes and actions, would then be assigned longer, less efficient longer word lengths.

In this paper, we continue to investigate evidence for extending and populating the communication spectrum, to enable understanding, detection, and categorisation. With such knowledge, any future decipherment capabilities will be enhanced, through development of resources such as an Affinity Matrix (Elliott, 2007; 2012), where models of constructs will be embedded for assisting categorisation.

Glossary

Entropy: In communication theory, entropy is interpreted as average uncertainty of choice, e.g., the average uncertainty as to what symbol the source will produce next or the average choice the source has to what symbol it will produce next. As values increase beyond H1 (First Order Entropy - non-dependent probability), symbols [events] are conditional upon prior symbols occurring: for example, H2 (second order) for the symbol before; H3 (third order) for the prior two symbols occurring etc. In language, the entropic value reduces due to its internal structure.

References

Elliott, J. & Baxter, S (2012) Constructing the Matrix: Acta Astronautica, Journal of the International Academy of Astronautics Elsevier Science Ltd, England.

Elliott, John. (2011), The Filtration of Inter-Galactic Objets Trouvés and the identification of the Lingua ex Machina hierarchy: Acta Astronautica, Vol 68 Nos 3-4, p399-405, Elsevier Press, England.

Elliott, J. (2007) A Post-Detection Decipherment Matrix: Acta Astronautica, Journal of the International Academy of Astronautics Elsevier Science Ltd, England, AA2853.

Elliott, J. (2003) 'Unsupervised learning of word classes using function word constraints in: proceedings of International Conference on Computer', Communication and Control Technologies (CCCT '03), Vol. V, pp. 143-148.

Elliott, J. (2002) 'Detecting Languageness', Proceedings of 6th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2002), vol. IX, pp. 323-328.

Freudenthal, H. (1960) *LINCOS, Design of a language for cosmic intercourse*. North Holland.

Ollongren, A. (2000) Large-Size Message Construction for ETI in: Proceedings of the 50th International Astronautical Congress, paper IAA-99-IAA.9.1.09, International Astronautical Federation, Paris.

Schulz, R., Wyeth, G., & Wiles, J. (2011) Lingodroids: socially grounding place names in privately grounded cognitive maps, Journal of Adaptive Behavior, 19 6: 409-424, ISBN: 1059712311421437

Shostak, Seth (2010) *What ET will look like and why should we care*: Acta Astronautica, v. 67, issue. 9-10, p. 1025-1029. ELSEVIER Press 2010 67.1025S